Statistical Inference Course Project

Alejandro

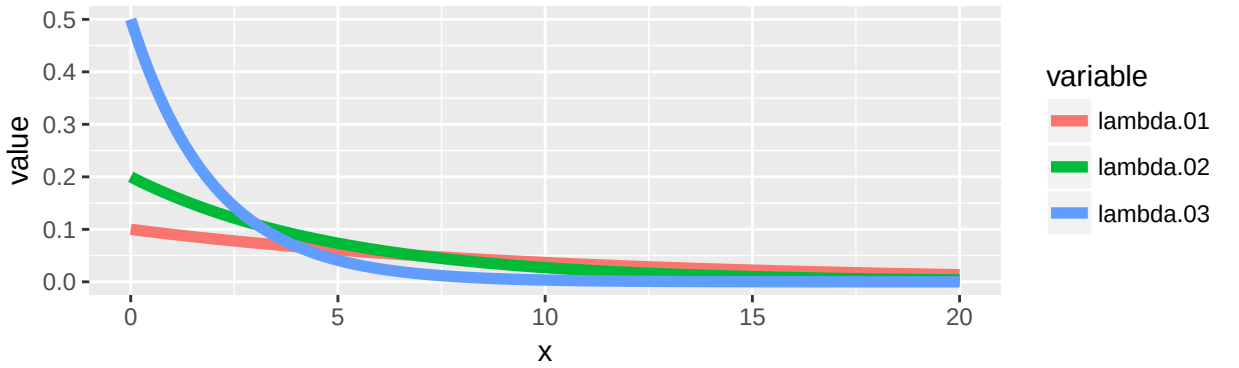
14 Januar 2018

Overview

In this Peer-Graded-Assignment the central limit theorem (CLT) and the ToothGrowth data will be analyzed. For the first part we will rely in the simulation power of R to empirically derive the CLT. In the second part the ToothGrowth data in the R datasets package will be analyzed using techniques learned in class such as confidence intervals and/or hypothesis tests.

Part 1: Simulation Exercise

We will use the exponential distribution and compare the mean of different simulations with the CLT. In the following graph we can see the exponential distribution plotted for different values of lambda.



CTL

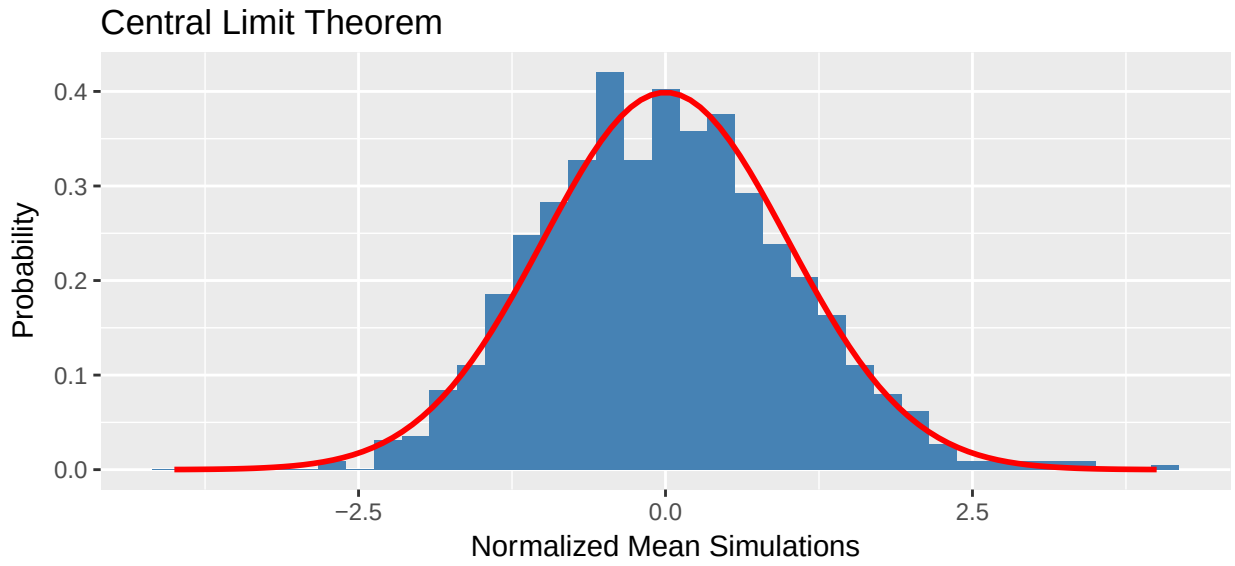
To analyze the CTL we will do a set generate 40 random numbers from the exponential distribution and compute the mean and standard deviation for this 40 random numbers. This will be done a total of 1000 times.

```
lambda <- 0.2
n <- 40
sims.mean <- NULL
sims.sd <- NULL
for(i in 1:1000){
    sims <- rexp(n,lambda)
    sims.mean <- c(sims.mean, mean(sims))
    sims.sd <- c(sims.sd , sd(sims))
}
```

In theory the mean and standard deviation for the exponential function should be 1/lambda. With a lambda of 0.2 the mean and standard should be 5 which is really close to what we obtained from the simulations of mean:5.01 and standard deviation: 4.93

In the following graph we see the the result of the simulations as a histogram as well as the normal distribution. It isnot a perfect fit but with a higher number of simulations the mean of the random numbers will look more and more like the normal distribution.

```
sims.mean <- (sims.mean - mean(sims.mean))/sd(sims.mean)
x <- seq(-4,4, length = 100)
hx <- dnorm(x)
g <- ggplot(data.frame(x = sims.mean), aes(x = x))
g + geom_histogram(aes(y = ..density.. ,fill = I("steelblue")), binwidth = density(sims.mean)$bw) +
    geom_line(data = data.frame(x=x,hx=hx),aes(x=x,y=hx), col = "red", size = 1) +
    labs(x = "Normalized Mean Simulations", y = "Probability", title ="Central Limit Theorem")
```



Basic Inferential Data Analysis

In this second part we will do statistical inference on the ToothGrowth data set from R. First we will take a look at the data set.

```
data("ToothGrowth")
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

We see that there is a total of 60 observations with 3 variables: len, supp, and dose. Variables len and dose are numerical variables while supp gibe us information about what supplement was use during the test being a factor variable with 2 levels. Using the summary function in R we can analyze the numeric variables and see their spread.

```
summary(ToothGrowth)
```

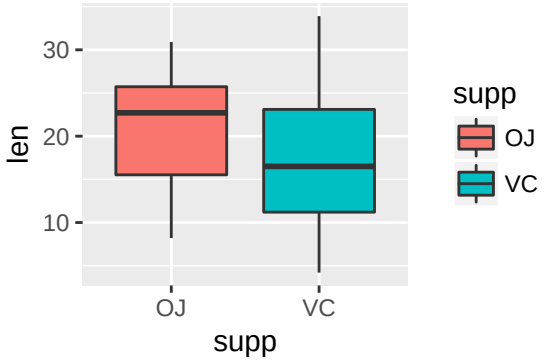
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

We can now see a boxplot for both dose to give us an better picture of the data. Furthermore we can see the effect that the dose has in the length.

```
aggregate(len ~ dose, ToothGrowth, mean)
```

```
##   dose    len
## 1  0.5 10.605
## 2  1.0 19.735
## 3  2.0 26.100
```

```
g <- ggplot(data = ToothGrowth, aes(x = supp, y = len, fill = supp))
g + geom_boxplot()
```



We see that the dose has a clear effect in the len. From the boxplot we can see that the supp changes the mean and the spread of the data. By using the t.test function in R we can compute the effect that both supplements have in the tooth growth.

```
t.output <- t.test(len ~ supp , data = ToothGrowth, paired = FALSE, var.equal = FALSE)
t.output$p.value
```

```
## [1] 0.06063451
```

With a p value of 0.06 we cannot reject the null hypothesis with a confidence of 95% that the different supplements have an effect in the tooth growth.

```
sub1 <- ToothGrowth[ToothGrowth$dose == 0.5 | ToothGrowth$dose == 1,]
sub2 <- ToothGrowth[ToothGrowth$dose == 1 | ToothGrowth$dose == 2,]

res1 <- t.test(len~dose, data = sub1, paired = FALSE, var.equal = FALSE)
res2 <- t.test(len~dose, data = sub2, paired = FALSE, var.equal = FALSE)
```

By looking at the pvalues: 0.0000001 and 0.0000191, we can reject the null hypothesis that the dose does not have an effect in the length of the tooth growth with a 95% confidence.

Conclussion

We can conclude that different doses lead to an increase in tooth growth, but we cannot say that different supplements lead to a difference in tooth growth.